DOCUMENT RESUME

ED 159 962 HE 010 411

AUTHOR	Eggert, John
TITLE	Developing Student Achievement Measures. Working Paper No. 3.
INSTITUTION	University of Mid-America, Lincoln, Nebr.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
PUB DATE	Dec 74
CONTRACT	NIE-G-75-001
NOTE	32p.; For related documents see HE 010 410 - HE 010 428
AVAILABLE FROM	Office of Public Affairs and Information Services, University of Mid-America, Box 82006, Lincoln, Nebraska 68501
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS	Accounting; Adult Education; *Course Evaluation; Educational Assessment; Educational Television; Higher Education; Institutional Research; Material Development; *Measurement Techniques; *Open Education; Psychology; Research Projects; *Student Evaluation; *Telecourses; Television Curriculum; *Test Construction; Test Items; Test Validity
IDENTIFIERS	State University of Nebraska; *University of Mid America NE

ABSTRACT

Student assessment procedures used by the University of Mid-America/State University of Nebraska (UMA/SUN) for its first two televised courses, Introductory Psychology and Accounting I, are presented in this report. Included are descriptions of various experiments in producing test items, a report on the construction of tests from the items, and a presentation of the results of the first tests in terms of learner achievement and item statistics. Consideration is given to the validity of UMA/SUN's tests and to the needs of the future. It is concluded that the process of specifying objectives precisely and meaningfully to inform both the student and the test designer of what is being taught works well and should be continued. (Author/LBH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

- ED159962

DEVELOPING STUDENT ACHIEVEMENT MEASURES

Working Paper No. 3

Office of
Research and Evaluation

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

John Eggert
University of Mid-America

December, 1974

HE 010 411

TABLE OF CONTENTS

Page

SOME EXPERIMENTS IN ITEM DEVELOPMENT

Accounting I - A First Experiment 1
Accounting I - A Second Approach 4
Accounting I - A Third Approach 8
Introductory Psychology - Item Development 10

CONSTRUCTING ITEMS INTO TESTS

Accounting I 15
Introductory Psychology 17

AN ANALYSIS OF THE INITIAL TEST RESULTS

Accounting I 19
Introductory Psychology 23

THOUGHTS ON THE VALIDITY OF UMA/SUN's TESTS 25

THOUGHTS FOR THE FUTURE 29

DEVELOPING STUDENT ACHIEVEMENT MEASURES

This is a report on the student assessment procedures utilized by UMA/SUN for its first two courses, Introductory Psychology and Accounting I. Included are descriptions of various experiments in producing test items, a report on the construction of tests from the items, and a presentation of the results of the first tests in terms of learner achievement and item statistics.

Some Experiments in Item Development

Accounting I - A First Experiment

During the week of May 6-10, 1974, a group of four people, consisting of three advanced doctoral students and one Assistant Professor, was gathered together for the purpose of developing achievement test items for the Accounting I course. This section describes the procedures of this experiment and comments upon the results obtained.

The Task. Participants at this first item writing session were to develop achievement test items for the Introduction and Unit A of the Accounting course. The Accounting course had been divided into units which are further divided into concepts. The Introduction contained six concepts and Unit A contained 22 concepts for a total of 28 concepts. Item writers were asked to study text chapters, the study guide and concept sheets in an effort to develop achievement measures.

The primary source for achievement development was the concept sheet. These sheets state an instructional task, prerequisite knowledge needed

by the learner to approach that task, and the content to be covered in a given concept. From those sheets, item writers were to construct achievement measures appropriate to that concept. The other materials mentioned, particularly the text book, were to be used as a reference.

It was hoped at least two drafts of each item would be developed during the week. These items would be finally reviewed for approval by the resident content specialist.

The Process. Item writers were first given an orientation to UMA/SUN. This orientation included special emphasis on the design and development of the Accounting course. Both the instructional designer and the resident content specialist associated with the Accounting course were present and offered their perceptions of the course as it now stands.

Each item writer was then assigned a total of seven concepts to deal with during the week. Each was given copies of the text chapter and the concept sheets and was to approach his task in any way he thought desirable.

Upon completing first draft items, each writer gave those items to an Accounting consultant, who was brought to Lincoln during the week to work with the item writers. That Accounting consultant reviewed each item, commented upon them and returned the items to the writers. The writers then revised the items according to comments from the content consultant. After the items had been revised, they were typed on index cards with one item per card. The cards were given to the resident content specialist for his review. Upon his recommendations, some items were rejected, others were recommended for revision and still others were accepted as is.

In all, 323 items were developed during the week. Upon review by the resident content specialist, 35 items were recommended for rejection,

267 items were recommended to stay as is, 21 items were recommended for minor revision.

Comments on the Process. The units under consideration in this first exercise were the beginning units of the Accounting course. Therefore, the content developed in those units was at a reasonably low level of complexity.

This factor is important in reviewing the results of the item writing exercise. Both Accounting specialists commented that the item writers had done a good job through approximately Concept 17 in Unit A. From that point on, the content became more complex, thus making it more difficult for non-content specialists to develop items. It appeared we would probably need to bring in Accounting specialists for the next round of item writing, and either give them the assistance of a measurement specialist, or have them work as a team with a measurement specialist. This, it was felt, might insure a better reflection of content depth than could be obtained from non-content specialists.

The items that emerged during the first round of item writing were, by and large, quite acceptable. The content specialist felt the items did capture the essence of the content being taught in the first units. However, as stated above, both content people remarked it would be increasingly difficult for a non-content specialist to grasp the subtle nuances of content that would be evident throughout the remainder of the course.

The total cost for the week of item writing amounted to approximately $2,600.00, or about $10.00 for each item acceptable as is. Given the number of items that were created and the quality of those items, we think it can be said that the experiment was successful.

Accounting I - A Second Approach

Following the conclusion and overall appraisal of UMA/SUN's initial effort at pinpointing effective means for developing course achievement measures, a second approach was planned and implemented.

A team of three persons experienced in Accounting, including one Assistant Professor and two advanced doctoral students, were assembled during the week of June 3-7, 1974. In addition, an Assistant Professor specializing in testing and measurement was brought in as a part of the team for the first two days. This team was charged, as was the first team, with the development of achievement test items for a stipulated portion of the Accounting I Course.

The Task. The team members in this second achievement measure development experiment were enlisted to develop appropriate test items for Units B through D of the Accounting I course. These three units, encompassing the material included in six chapters of the text (two chapters per unit), included 37 concepts. The team members were provided with concept sheets and text chapters for the units assigned. In addition, they had access to all the course materials used by the first team and to the items developed by it.

The team was informed that it was anticipated that more than one item would be developed for each concept; that all concepts would have items developed relative to them; and that each item developed would go through at least two drafts during the week, after which the resident content specialist would review the items for final approval.

The Process. As with the first team, a rather thorough orientation to UMA/SUN and to the Accounting course was the initial undertaking

with this second team. It was originally planned that the following procedures were to be followed by the team. Each of the three accounting specialists was assigned one unit arbitrarily for item development. As items were written, they were to be reviewed by the measurement specialist during the first two days in order to ensure compliance with correct item writing standards. It was assumed this step would be a necessary provision only during the initial stage of the team's work. The items, when approved by the measurement specialist, were then to undergo initial review by the resident content specialist, to insure adequacy of the specific content of each item. When a team member completed writing items for all the concepts in the unit, a second team member, again assigned arbitrarily, was to go back over that unit and develop items considered necessary for adequate coverage of the concepts. This 'two vantage points' approach would provide, it was felt, sufficient items for each concept.

The unexpected illness of the resident content specialist necessitated a significant change in the planned procedures for the team. It was decided the team as a whole would serve as the "reviewing content specialist." Each team member write the items for his assigned unit as planned, and these items were then reviewed by the measurement specialist during the first two days (and thereafter by the resident Research Associate). When a team member completed the items for his unit, they were turned over to a second team member for review as to the adequacy of content and for the development of additional items if necessary. The team members were located in the same room (being from Lincoln, they had the option of working elsewhere) and they sought to resolve differences regarding item content

through open discussion. Following this internal review by the team, the items were typed on individual index cards which were reviewed by the resident content specialist for final approval.

A total of more than 250 items were developed during the week. Of this total, 40-50 were considered to be of too low quality to be useable, and 20-30 were considered to be of useable quality with some modifications. The remaining items were considered to be of high enough quality to be considered for use in the actual test.

Evaluation of the Approach Used and the Process Followed. This second approach to achievement measure development has seemed to provide some definite advantages over the first approach employed. In the first place, the team members were well-acquainted with the content involved and thus did not require the familiarization period needed by the first team. They were experienced in writing test items over the same or, at least, very similar accounting material. Secondly, although the total number of test items generated was less than with the first team, the units covered by the second team were considerably more involved and the items developed were commensurately more detailed and complex. Many of these items were, as before, relatively short objective type items, but even here they were typically of greater complexity than similar items relating to the earlier units. Thirdly, the team members were already well-acquainted with each other, and were able to work together smoothly and efficiently from the beginning.

In the light of the difficulty encountered during the first team writing effort in having measurement specialists who were previously unacquainted with the content develop the items, this approach does seem at first blush

to be a more effective one. The only problem encountered was the time the members of the team required at the outset to familiarize themselves with the sequencing of the material, the procedures to be followed in the development of items, and the review of what had been already done by the first team. The team members themselves suggested that, if possible, in the future it would be more efficient to have the same persons develop all the items for the course. It was further suggested, however, that if this is done, the task should be aplit so such persons spend no more than a week at a time on concentrated item writing. It has become apparent to all involved that the efficiency of writing test items decreased markedly after several full days of continuous writing.

This particular team displayed more item writing sophistication than was originally assumed. As a result, there was some question on the part of team members (including the measurement specialist) as to whether the procedure followed in using the measurement specialist was the most effective one. It was suggested it might be more efficient, at least under these circumstances, to have the measurement specialist review the items toward the end of the item writing period rather than at the beginning. If this were done, it should perhaps be coupled with two other provisions: 1) a resident staff member in the Office of Research and Evaluation should review all items as they are developed in order to pinpoint any general deficiencies; and 2) the item writing team should have selected materials which provide guidelines for correct item writing.

Although the items generated by this team have not yet been subjected to any comprehensive analysis, they appear to be of a reasonably high quality. Moreover, the degree of learning which these items seek to measure seems

to be commensurate with the degree of learning sought from the instructional material.

The costs for this second venture in item writing for the Accounting I course were significantly less than in the first experiment because of the reduced travel and per diem costs for the team members. The cost per usable item under these conditions was about $5.00. The same conditions (availability of competent content specialists locally) will not always exist with other UMA/SUN courses but, nonetheless, it appears at this juncture that the quality of items produced, the number of items developed and the costs involved worked together to bring about a successful achievement measurement development venture.

Accounting I - A Third Approach

Having experimented first with a team of measurement specialists working under the guidance of content experts, and then with a team of content experts working under the relatively direct guidance of a measurement person, a third approach to item development was tried. The content experts wrote relatively independently, with their work being checked only by other content experts before being delivered to UMA/SUN's resident content and measurement specialists for consideration for use in the actual test.

The Process. The items which were to be written during this final session were to cover Units E and F of Accounting I, the last two units of the course. The team consisted of the same three persons experienced in accounting as had participated in the second session, plus the resident measurement specialist of the Office of Research and Evaluation. The primary difference in format was that all of the writing was done entirely in the homes or offices of the content persons, with all contact

between team members done through the U.S. Mail. As items were written based on the usual concept sheets and textual material, they were forwarded by mail to another team member who reviewed them for content accuracy and sent them to UMA/SUN's measurement specialist. All other details were essentially the same as those used in the second approach.

Evaluation of the Third Approach. The fruits of familiarization with the task, with the content and with the other team members were evident during this third session. The team required little additional direction. They produced slightly more items (about 300) than were produced in the previous session, with about the same proportion of items acceptable as is.

There are some hazards to be avoided in this approach, however, if it is to be adopted. It is important the progress of the writers be monitored, to insure the work is done on time. The writers on this team presented no problem in this regard, but the temptation to delay the work in the informal environment of a writer's home or office is probably a real one.

It is also important the writer's work be monitored for quality and the quantity is specified. Again, this group posed no problems, because they had previous experience with UMA/SUN, and they were well-qualified. However, as new content persons are selected to write items for various courses, it will be necessary to closely monitor the quality of the writing during the early stages of the process, lest we be surprised with unexpected incompetence in the face of a deadline.

The problems with this third approach, then, are problems of communication. If the team is well-qualified, and thoroughly familiarized with UMA/SUN's needs, there should be no problem. If there is uncertainty about the team's qualifications, it is important their work be carefully

scrutinized during the early states of item development. In general, the items produced in this session appear to be of higher quality than those from the first session, and as high a quality as those from the second session. It is not possible to give a percentage estimate of numbers of unacceptable items, however, as the decision process consisted of a gradual narrowing down of the item pool by the measurement person and by the resident content specialist, who became more particular at each stage of the selection process, until the final selection of items had been carefully scrutinized by both content and measurement people a number of times. The items were of such quality that the final decisions on item selection were generally between an adequate item and a very adequate item. This was also due to the fact that we have a relatively large item pool from which to choose.

The cost per item was relatively low, about $3.00 per item. This was due to the absence of travel costs, the previous training and experience of the team members, and the fact that the measurement specialist was on staff as opposed to being brought in as a consultant. In the future, for courses which have well-specified content, and for which there is a clear understanding of what is evidence of comprehension of that content, UMA/SUN would do well to follow this model of item development, possibly in conjunction with the second approach as an introduction.

Introductory Psychology - Item Development

During the week of September 9, 1974, a number of persons were brought together to create an item pool to be used for the construction of tests for the Introductory Psychology course. The team consisted of two University of Chicago doctoral students in measurement who were also knowledgeable in psychology, and three University of Nebraska at Lincoln (UNL)

graduate psychology students, each of whom had some experience teaching general psychology and/or writing psychology test items. These five students and a University of Nebraska at Omaha (UNO) professor actually wrote the items, while another UNO professor of psychology and UNL professor of psychology participated as reviewers and occasionally as item writers.

The Task. The first morning was spent orienting the team to SUN, the Introductory Psychology course and the task at hand. Some time was also spent in discussing item writing, in terms of examples collected for that purpose.

The writers were presented with concept sheets, representing the terms and phrases thought to be important by the course content specialist, one sheet for each of 15 units. The writers took turns choosing units on which to write items, until there were two persons on each unit. They were instructed to write one item for each concept. Through discussion, they were particularly made aware we wanted a minimum of "fact or definition" items, and to emphasize the higher-order behaviors of analysis, comprehension and synthesis. We requested they write questions requiring the students to extrapolate from the text, and to apply what they learned to new situations. We also gave the item writers a list of objectives showing that the student was held responsible for defining terms, for relating them to one another where appropriate and, when given an example, for labeling the concept it illustrated.

The Process. The procedure was for the writers to write the items, for the items to be typed by the secretaries, and then to be reviewed by each of the reviewers in terms of difficulty and content coverage.

This procedure went on for the remainder of the week, with half-hour seminars everyday before lunch to discuss the procedure. In addition to writing the items, each writer was asked to write feedback to the student to increase the instructional value of the test taking experience.

On the experience that item writing can be a draining process, and that continued writing results in mental fatigue, and thus lower productivity, a couple of strategies were used. One strategy was to provide a "game room" where a soccer-board game and a carom game were available. The writers were also encouraged to leave the building for frisbee, or whatever recreation they could find. Also a trip to an exotic restaurant for chocolate pancakes was placed in the schedule, as was a group picnic.

Comments on the Process. During and at the end of the session, the item writing participants were asked to give their perceptions of the procedure. A summary follows:

The introduction to SUN was thought to be too lengthy (half hour on SUN, slightly less on Introductory Psychology). The discussion on item writing was viewed with mixed emotions. The psychology students thought it was useful, but that 45 minutes was too long.

The concept sheets were viewed as necessary, though there were many complaints that they were inadequate. It was felt some concepts were perhaps overemphasized, some underemphasized, and many were trivial. It would probably be a good idea if a means of weighting the concepts could be devised, in order to give the important concepts more emphasis.

A means of stating objectives more explicitly also needs to be developed, so it is clear to the writers when only definitions are desired, or when high-order questions are necessary. Many of the concepts demanded the writing of "fact" or "definition" items.

It seemed to be a good idea to let the writers choose the units on which to write items, and it was fortunate they came from a wide variety of psychological backgrounds. It was felt by the writers, and perceived by the reviewers, that providing feedback to the writers was counterproductive, too much time was then spent on modifying, and sometimes arguing, with the reviewer. However, some writers felt very strongly they should see what the reviewers said.

The game room was never used, though the frisbee was. The trip out for lunch was good for all, and another trip or two out would have been helpful.

The morning item writing sessions were felt to be most productive. Some thought a long morning with no writing in the afternoon might be a good idea. Another suggestion was that the morning be spent writing items, and the afternoon writing feedback and studying for the next day. We generally started about 8:30, broke for discussion at 11:45, then ate, and started again about 1:15, working until 5:00.

It was suggested, and generally agreed, that it would be best if the writers were paired together on groups of chapters, so they could work more efficiently.

Another point made was that the reviewers should be hired to come in after the first day or two of writing, and to remain an additional day or two after the writers completed their task.

During the week, six item writers wrote 520 items. Of these, about 120 were rejected by either content specialists or the measurement specialist. Though we desired feedback on half of the items, only about one fourth of the items had any adequate feedback. The item writers average about three

items per hour, with quite a bit of variability in quantity and quality.

Conclusions about the Item Development Process. In general, we are quite satisfied with the items developed through the above processes. In terms of quality, it appears that while measurement specialists can provide some initial insight and direction to a group of content specialists, and while their expertise is necessary in the process of transforming the raw items into a test, the wisdom of their use as actual item writers is doubtful.

We are also well satisfied with the coverage of the material, due primarily to the rather detailed concept sheets which were provided to the writers. Such lists of very specific concepts, or objectives, seem to be indispensible to the item writing task.

In terms of guidance required by item writers, it appears if the writers are thoroughly familiar with the content and have been explicitely told what is needed, there is little need for a close monitoring of the actual item writing process. By the same token, pains must be taken to insure that the details of what is needed are in fact explicitly understood.

Clearly the single most important variable in terms of cost per item is travel and accomodations for writers. The implication of this is that when writers and consultants need to be hired, the Lincoln area should be searched before turning to out-of-state sources; or, alternatively where feasible, communication with such writers be via telephone and U.S. Mail.

Constructing Items Into Tests

Accounting I - Test Construction

Once the raw items have been developed, the process of selecting and organizing them into test form can begin.

The first step in this process is to sort the raw items according to the concepts they were designed to cover. Each item is labeled according to concept as it is written. The resident content specialist is then consulted for advice on how he would like the concepts to be weighted. Some concepts include a large amount of detail and must be sampled extensively, while other concepts may require only a single item. At the same time the content specialist advises the measurement person on such issues as preferred format, degree of variety desired, length of test, total number of items, etc.

The next step consists of the measurement specialist selecting 2-3 times as many items as are needed for each concept. This initial selection is done somewhat superficially, with obviously bad items being rejected, and with an attempt made to obtain a variety of form, style and detail of content within specific objectives.

The next step is also performed by the measurement person, who looks more closely at each item and each alternative, searching for inconsistencies, obviously wrong alternatives, give-away answers, grammatical problems and general item quality. The items surviving this editing process, still substantially more than are required for the actual test, are then given to the content specialist.

The content specialist performs essentially the same operations on the items, selecting the best item or items within each concept, the main difference being that his criteria are primarily related to content, not to psychometric considerations. The prime criterion of both the content person and the measurement specialist is characterized by the question, "Does a person who gets this item correct probably understand the concept?"

Usually the content person will decide there is a need for an additional item or two, in which case he will locate an appropriate question in the remaining pool. Additionally, he will certainly make a number of corrections and revisions of the chosen items as he sees fit. Once he feels he has a test with which he is satisfied, he returns it to the measurement specialist who once more checks it over. The test is then typed for administration to a prototype test group.

The prototype test group, in the case of Accounting I, consists of six persons who were paid to take the first course. They take each test first as a pre-test, and then as a post-test. They are asked to indicate any inconsistencies, misspellings, or any other errors they might notice. The frequencies of response to each alternative are examined for any suggestions of possible problems. Indicators of problems might include instances in which the majority of persons select a correct answer on a pre-test, or the majority select a particular incorrect alternative on a post-test. Those items with suspicious response patterns are then thoroughly scrutinized and discussed by the content person, the course evaluator and the measurement person. Often corrections are made, but sometimes it is decided that there is nothing wrong with the item, and it is left as is.

Finally, the test is retyped on a stencil in final form and is handed over to the SUN delivery system for administration to learners, after which it will once again be analyzed, reviewed and revised as necessary.

Introductory Psychology - Test Construction

The test development process used for Introductory Psychology was essentially the same as that used in Accounting I. However, the review

and editing procedure required a substantially greater degree of resources, since more tests and longer tests were required.

Each item was first examined in terms of the usual general requirements of clarity, grammar and consistent format. The items were then sorted according to unit arrangement and by concept within each unit. Items were then sampled within units according to a weighting formula determined by the content person with more items selected from one or another unit depending on its emphasis in the course. Enough items were sampled to allow for the construction of two tests for each of the five test periods--the real test and a parallel trial test. The trial test was to serve as practice in test taking, as a means of self-assessment and as an additional indication of the objectives of the course.

The First Drafts. An examination of the resulting drafts of the tests revealed some inadequacies in the sampling procedure. It appeared some concepts were over-sampled, and some under-sampled. Moreover, in some cases, it would appear to the learner, upon taking the trial test and then the real test, that she could determine (by deductive analysis) which specific concepts would be selected for the latter by examining the former. Such a cue to the student would certainly be misleading.

Part of the reason for the ineffective sampling was no doubt due to chance. In reality, particularly on the trial test, it is desirable to have a sampling of items that is more uniform than chance would provide, lest a duplication of a concept mislead the learner in terms of its relative importance. Also, we want to let the learner check herself on as many different concepts as possible in a trial test of given length.

Another reason for the ineffectiveness of the sampling procedures was the nature of the items. It was rare that any one item really covered

only one concept. More often two or more concepts are covered by a single item. Because some concepts are more easily duplicated than others, they are likely to be over-sampled.

A final reason for the inadequacy of the sampling procedure was we simply did not have enough good items for all of the concepts. Though initially at least one item was written for each concept, by the time they were edited at various stages of the test development process, many items were dropped and some concepts were left uncovered.

Revision of the First Drafts. Our response to the sampling of items for the ten tests (five trial tests and five "real" tests) was based on intuition and common sense. If some concepts appeared to be over-sampled in a trial test, but not the real test, or vice versa, we shifted items from one test to the other. If the two tests looked too similar to each other in terms of specific concepts covered, we dropped items from one, and added different items. If it looked like there were some large holes in either test, with important concepts being neglected, we made plans to obtain additional items.

It also became obvious, as the ten tests took form, that the student feedback on the trial tests was in general inadequate. There were wide variations in style and form, primarily due to the variation in interest, skill and priorities of the item writers. Feedback for many of the items was limited to a brief page reference, while for others it was several sentences long.

In order to develop these rough drafts into polished tests, two writer-editors were rehired from the original team of item writers. One was the doctoral student-instructor who had shown himself to be sensitive to the task, and the other was the UNL professor who had proven to be a dedicated critiquer.

The doctoral student was given the drafts, and told to write, rewrite and edit them into polished form according to our specifications, and according to the comments of previous reviewers. He was asked to pay close attention to the feedback on both sets of items (those for the real tests as well as the trial tests) to insure the items were in fact covered in the text, and as a further check on their correctness. It was felt the feedback might eventually be used in conjunction with the real tests as well.

As the tests were edited and rewritten in this manner, they were forwarded to the UNL professor, who performed essentially the same task, only with particular attention being paid to coverage of content. A number of details of form and style were also corrected at this stage.

The next stage of test development was a review by the resident measurement specialist and the resident content specialist. After a final revision, the tests were stenciled and duplicated.

An Analysis of the Initial Test Results

Accounting I - Initial Test Results

The Mechanics of Administration. Sometime before the first students are expected to take the test, all of the students are contacted by mail and advised on the alternative methods of taking the test. The students have three options: They can come into the learning center at one of several prescribed times; they can call, write or visit a learning center to make an appointment to take the test at any time; or they can make arrangements through the learning center to take the test at a

library in their county seat, the librarians of which have previously consented to participate.

The learners are allowed 90 minutes to finish the exam, though most finish in less than an hour. The tests are proctored by the librarians or the learning center staff. The learner may use no texts, notes or references. The answers to the multiple choice items are entered by the learner onto an IBM scoring sheet. The actual test is destroyed by the proctor, and the answer sheet is mailed in to the faculty office at SUN Central (Lincoln). The sheets are Xeroxed, with the original sent to Research and Evaluation, where they are analyzed as soon as there are enough to provide a quick statistical check on the first use of the test. The copies are corrected and graded by hand. Test scores and grades are sent to the students after having been determined by the faculty. The score sheet copies are filed for faculty use in follow-up phone calls, learning center visits and other student contacts.

How the Learners Fared. As of January 15, 1975, or about 60 days after the day a learner could reasonably be expected to take the first Accounting test, 180 or approximately 67% of the students taking Accounting I for credit, have actually taken the test. In keeping with SUN's philosophy of allowing students to proceed at their own pace, no determined effort is being made to systematically prod those students who are behind. However, it is also recognized that some students may need and even desire a little external prodding, and a general letter to all students was sent out, over the instructors signature, congratulating those students who attempted the first test, noting that the results weren't bad, and gently urging those who had not yet taken the exam not to put it off unless they

"really felt it was necessary." Plans are being made to evaluate the situation again after the holidays, at which time more direct action may be considered.

The scores on the 28 item test have been analyzed in 3 batches. The mean score for the group of 56 students who took the test during the 10-day period after it was first offered was approximately 25, with a range 17 to 28.

The mean score of the 59 students who took the test in the following 15-day period was approximately 22, with a range of 11 to 28.

The mean score of the 65 students who have taken the test through January 15, 1975, is also approximately 22 with a range of 11 to 28. The overall mean for all 180 students is 22.5.

The course professor assigned grades according to the following schedule: 26 (92%) and above is A; 23 (82%) and above is a B; 19 (67%) and above is a C; 16 (57%) and above is a D; and below 16 is an F. For those electing the pass-no credit option, 16 is also the cut off point. In this first sample of 180 students, on the first test, there were 38 A's, 69 B's, 51 C's, 10 D's, and 12 no credits. The grading schedule was arrived at by the course professor, based on his own experience teaching Accounting I on the University of Nebraska-Lincoln campus. Further certification of the grading schedule is planned.

Item Statistics. The Accounting I tests are essentially Criterion Referenced Tests, that is, each item is directly linked to a domain of content specified in a statement of course objectives, and success is determined by attainment or non-attainment of objectives rather than by relative rank in a class. Therefore, classical test analysis statistics are not entirely appropriate. This is particularly true when a number of

items are answered correctly by a very large proportion of the students, as was the case in this first exam.

Those items that were answered correctly by a very large majority of the students have been carefully re-examined by the measurement specialist and the course instructor to insure that such items have not been "give-aways", but truly reflect the attainment of the relevant objective. While about 1/3 of the items were answered correctly by over 90% of the students, no such "give-away" items have been found.

With such low score and item variability, classical discrimination indices such as the point-biserial correlation become meaningless. Discrimination was checked in two ways, however. First, the students were divided into groups of high scorers, medium scorers, and low scorers. Then, for each item, the percentage of each group answering it correctly was calculated. If the item did discriminate between students of various levels of ability, one would expect the higher level group to have the highest percentage correct on each individual item, and the middle group to have the second highest percent correct. In fact, this was the case with every item, excepting those items which were answered correctly by all members of the top two or of all three groups, and excepting a couple of statistically insignificant differences.

Another sort of discrimination index is the mean total score of all persons choosing the correct answer on any one item. If choosing the correct response to the item is positively related to ability, it should be true, in general, that the mean score of those persons choosing the correct answer for any one item should be higher than the mean score of persons choosing any other alternative for that item. This was universally

the case, except for a couple of statistically insignificant differences.

Finally, a split-half reliability coefficient was calculated on each of the first two batches of tests, the prime difference between the two batches being the range or variation of the scores (the variance of the second test was three times that of the first). For the first batch, the reliability coefficient was .547, not at all unreasonable given the low variation in scores. The reliability coefficient of the second batch was .824, a very satisfactory coefficient in classical terms, even without a restricted range.

In fact, however, while the reliability coefficient does provide worthwhile information in a given context, the fallacy of focusing on such an indicator of test quality is well-illustrated here. The same test administered under almost identical conditions, to nominally identical populations, has two very different reliability coefficients--one quite mediocre and one quite satisfactory!

What is important is that each item directly tests competence in a given objective, and that all course objectives are adequately sampled. The above statistics help us ascertain whether or not this goal was achieved, but the reliance has been and must be on the careful process of linking items to objectives.

Introductory Psychology Test Results

The tests were administered in the same manner as the Accounting I tests, with the exception being the pre-distribution of the trial tests and accompanying feedback.

The trial tests looked essentially like the real tests, except the cover sheet explained how the trial test might be used and interpreted.

It was suggested the student study until he felt ready to take a test and then attempt the trial test under conditions similar to the real test (e.g., without the aid of books or study materials). After taking the trial test, it was recommended the learner should check answers and read the feedback to the questions she is uncertain about. Based on her experience with the trial test, she should then either study more, as suggested by the feedback, or attempt to take the real test.

The correct answers to the trial tests, each with feedback of two or three sentences of explanation and page references, were provided in a separate booklet that had been stapled on all four corners. The student was instructed to snip off three of the corners to separate the pages so she could check her answers and read the comments. This method was to encourage the learner to treat the trial test as a real test, to help insure that the scores would be comparable.

How the Learners Fared. As of December 9, 1974, 104 or 38% of the Introductory Psychology students had taken the first 48-item exam. The scores ranged from 11 to 45 correct, with a mean of 36. The grading scale, based on the course professor's judgment, his experience with similar courses, and the opinions and practices of other professors of introductory psychology courses, is as follows: 42 (88%) and above is an A, 38 (79%) and above is a B, 33 (69%) and above is a C, 27 (56%) and above is a D, and below 27 is "No-Credit". In this first sample of students there were 19 A's, 30 B's, 28 C's, 19 D's, and 8 No-Credits.

Item Statistics. Overall, the item statistics look fairly good. The total test had a Kuder-Richardson Formula #20 reliability coefficient of .831, which we consider to be quite adequate, given the reservations about the utility and validity of such statistics noted in the

Accounting I report. The indices of discrimination discussed in that report suggest that three items will need to be revised. Inspection by the course professor revealed the material for one of these items was not well-covered in the text, and the two other items were ambiguously written. The students who have already taken the test will have their scores adjusted accordingly.

The classical item statistics were more applicable to this test than they were to the Accounting I test, because of the greater variability of scores and the higher level of difficulty of the items. No items were answered correctly by all. Fifteen of the 48 items were answered correctly by more than 90% of the students, and another 15 were answered correctly by less than 70% of the group. While this is less than ideal if we really do want all of our students to know all of the concepts, it does allow us to use the classical indices with some appropriateness.

It is still too early to judge the worth of the trial tests. Informal feedback via the learning center coordinators is that the trial tests have been well-received and appreciated. We will be asking more questions about these trial tests in future evaluations.

Thoughts on the Validity of UMA/SUN's Tests

The question is often asked, "Are our tests valid?" which asks in technical jargon the simple-minded question "Do they mean anything?" Of course, the answer is "Yes, and No". The following general discussion reflects on experience both with Accounting I and Introductory Psychology.

If the purpose of our tests is to determine the extent to which the material thought important by the content specialists and/or the academic

community is learned by the student, then the crucial concern is with content validity. This means, "Does the test adequately sample all of the material considered important?" In UMA/SUN, we deal with this question by listing all of the objectives of each course as it is produced, writing test items to cover each of these objectives and sampling these items to be as inclusive of content as possible. The items thus written are reviewed in terms of adequate coverage of content and general item quality by one or two outside content specialists, our resident content specialist and our measurement specialist. In addition, the sample of items chosen for the test is given a final check by the content specialist and, in the case of Accounting, are administered to a prototype sample of students as a further check. By this process, and to the extent the objectives are valid, we can feel confident in claiming our tests have content validity. That is, they are an adequate reflection of student competence in terms of the stated objectives.

There are some areas of concern in this process, however. The first is the implicit, sometimes explicit, statement of objectives by the content people. Before these objectives are incorporated into the test construction process, we have to make sure we truly want them to be UMA/SUN's objectives. If we do not want them to be UMA/SUN's objectives, then they should not be incorporated in the test and the test may not have content validity in terms of this sector of the academic community. On the other hand, if we want content validity in terms of content people's conception of what is important content (or anybody else's conception, for that matter), then we must make sure our final objectives reflect the views of these people. It should be noted this aspect of UMA/SUN's test construction

so far has not been carefully controlled. In the case of Psychology, these objectives had to be contrived ad hoc, almost post production. This is risky business. In this case, we are assured of content validity in terms of the stated objectives but we must be aware of the fact that these objectives were derived from the course, not directly from the academic community, nor from a consideration of the purposes of UMA/SUN.

Probably the next most relevant type of validity with which UMA/SUN should be concerned is concurrent validity, that is, with the question "Does this test measure the same thing as some other test?" This is the empirical counterpart of content validity and similar cautions must be observed. The most important is a check to make sure we want out test to measure what the other test measures.

For those who define academic credibility as proof our students are capable of passing the same tests as other students, concurrent validity is important. A number of questions must be asked, however, such as "Are these other tests valid?" "Do they measure what we, or the academic community, feel is important?" "Is the content of the tests, we well as the courses for which the tests are designed, the same?" Given that these questions are answered affirmatively, then it makes some sense to administer UMA/SUN's test to these other students, or the other tests to our students, and note how well these two tests correlate. This could be done by paying students to take both exams, constructing tests composed of half of one test and half of the other, or by similar methods. An alternative to using an existing test designed for only one course would be to use a standardized test accepted by the academic community.

Presently we are minimizing the use of such procedures. We could have students in similar courses offered by other institutions take some

of our tests. But it is probably true there are no other courses that stress what, and only what, UMA/SUN's present two courses stress. For this reason, the best information we could hope for is some gross indication of item difficulty and clarity. We have to rely heavily on our test development process, including the usual post hoc test analysis procedures that check internal consistency, to back up our claims of validity.

The type of validity tacitly assumed to be present by most educators, but is probably the most shakey, is construct validity. This is represented by the question, "Are we measuring what we think we are measuring" or "Is the meaning we give to the scores the meaning the scores should have?" Academicians are fond of assuring one another they know their tests don't measure what is really important, but yet, they generally continue to give grades based entirely on the test. We're not ready to suggest there is any real alternative to this, but particularly in an institution such as UMA/SUN, that stresses innovation in serving heretofore unserved needs of students, it would seem foolish to be tricked into looking at test scores, often traditional by design as the sole indicators of success or lack of it.

To summarize: The answer to the question "Are UMA/SUN's tests valid?" is "We try." We strive to develop objectives that adequately reflect the intent of the course, and have done a fair job in doing so. We create items reflecting these objectives under close supervision of content specialists. These items are sampled comprehensively to make up a test that is once more reviewed by content specialists, which is then prototype tested for obvious faults. This does a fair job of insuring content validity. We also have plans for trying other tests on our students, and our

tests on other students. These plans are not yet solidified. This will be our test of concurrent validity. We should always be wary of construct validity.

Thoughts for the Future

Certain aspects of the learner-centered information system at UMA/SUN, in general, seem to be in good order. The process of specifying objectives precisely and meaningfully to inform both the student and the test designer of what is being taught works well and should be continued. The content-expert team approach to item writing will be continued, probably with the initial sessions being held at UMA/SUN, but with much of the work being done at the team members' homes and offices. Parallel Accounting I tests are being constructed to allow students to retake a test, and other forms of information systems are being contemplated, including computer-managed and computer-assisted instruction. The parallel self-tests with annotated answers used in Introductory Psychology also appear to be useful. The test construction and review process seems to be in good shape, as is the test distribution/administration system. As the course progresses, we will remain alert to the possible need for diagnostic and remedial testing. We will continue to closely monitor the validity of our tests, primarily through a strong emphasis on referencing items to well-developed objectives and through expert review of the test development process at all stages. Finally, we will continue to ask the question "Whom are we serving and why?"